Research Project Questions

DATA5207: Data Analysis in the social sciences

NAME

December 16, 2024

Contents

1 Introduction

The monitoring of health statistics forms creates a strong foundation for the tracking of the general population health and guides policy formulation around matters health, as Krieger et al., (1997) point out. The collection of health statistics extended to socio-economic aspects ever since Massachusetts became the first state in 1842 to formally adopt it. Various studies have been undertaken to map out the relation between health metrics and socio-economic aspects, including (Braveman et al., 2010) checking on education, income levels and race and found that the least educated and lower income groupings were often associated with lower health status.

Adler & Newman, (2002) review socio-economic status, and the associated impact on health. Access to education eases the promotion of health information to individuals, access to income provides avenue for better nutrition, access to education and housing needs. Occupation on the other hand provides access to health related benefits, while also providing exposures to physical injuries for certain occupations.

The main objective of the analysis undertaken herein is pointing out the predictors for better health outcomes at the county level in the United States. The relationship existing between various variables is assessed and how they generally influence the number of deaths among United States of America county residents under age 75 per 100,000 population (age-adjusted).

The dependent variable is the premature age-adjusted mortality variable, while the predictor variables are selected based on previous research (Cheng & Kindig, 2012). The predictor will include variables with information on income and income inequality, population, demography aspects including race, health care costs and associated level of access, and numbers on primary care providers, preventable hospitalizations, high school graduation rates and college education levels, the percentage of single-parent households, and children living in poverty guidelines, and finally percentages of adult obesity and smoking.

2 Methods

A multiple linear regression model is used to investigate the existing relation between the premature age-adjusted mortality variable and its explanatory variables. The data of different health outcomes at the level of US counties, is obtained from the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute.

2.1 Data Information

The data contains 3142 observations and 20 features.

2.1.1 Measures

Dependent Variables: `Premature age-adjusted mortality`

Independent Variables:

- Median household income: Small Area Income and Poverty Estimates
- Income inequality: Ratio of household income at the 80th percentile to income at the 20th percentile
- Population
- % Non-Hispanic African American:
- % American Indian and Alaskan Native:
- % Asian:
- % Native Hawaiian/Other Pacific Islander:
- % Hispanic:
- % Non-Hispanic white:
- Uninsured adults:
- Health care costs
- Other primary care providers
- Preventable hospital stays: Number of hospital stays for ambulatory-care sensitive conditions per 1,000 Medicare enrollees
- High school graduation: Percentage of ninth-grade cohort that graduates in four years
- Some college: Percentage of adults ages 25-44 with some post-secondary education
- Children in single-parent households: Percentage of children that live in a household headed by single parent
- Children in poverty: Percentage of children under age 18 in poverty
- Adult smoking: Percentage of adults who are current smokers
- Adult obesity: Percentage of adults that report a BMI of 30 or more

2.2 Descriptive Statistics

Summary statistics of the data is as below:

Table 1: Table continues below

age_adjusted_mortality	household_income	population
Min. : 133.0	Min. : 22045	Min. : 88
1st Qu.: 323.1	1st Qu.: 41072	1st Qu.: 10976
Median : 388.3	Median : 47589	Median : 25771
Mean : 401.3	Mean : 49522	Mean : 102841
3rd Qu.: 470.9	3rd Qu.: 55308	3rd Qu.: 67490
Max. :1142.6	Max. :134609	Max. :10137915
NA's :65	NA's :1	NA

Table 2: Table continues below

percent_african_american	percent_american_indian_alaskan_native
Min. : 0.0000	Min. : 0.0000
1st Qu.: 0.6662	1st Qu.: 0.3609
Median : 2.1734	Median : 0.6105
Mean : 8.9426	Mean : 2.3093
3rd Qu.:10.0817	3rd Qu.: 1.3053
Max. :85.1516	Max. :93.0675
NA	NA

Table 3: Table continues below

percent_asian	percent_native_hawaiian_other_pacific_islander
Min. : 0.0000	Min. : 0.00000
1st Qu.: 0.4379	1st Qu.: 0.02824
Median : 0.6959	Median : 0.05731
Mean : 1.4856	Mean : 0.13499
3rd Qu.: 1.3522	3rd Qu.: 0.11214
Max. :44.2658	Max. :50.00000
NA	NA

Table 4: Table continues below

percent_hispanic	percent_non_hispanic_white	percent_uninsured_51
Min. : 0.5018	Min. : 2.812	Min. : 2.616
1st Qu.: 2.2071	1st Qu.:64.904	1st Qu.: 9.109
Median : 4.0995	Median :83.989	Median :13.567
Mean : 9.2896	Mean :76.584	Mean :14.289

percent_hispanic	percent_non_hispanic_white	percent_uninsured_51
3rd Qu.: 9.4262	3rd Qu.:92.712	3rd Qu.:18.315
Max. :96.2540	Max. :97.977	Max. :43.395
NA	NA	NA's :1

Table 5: Table continues below

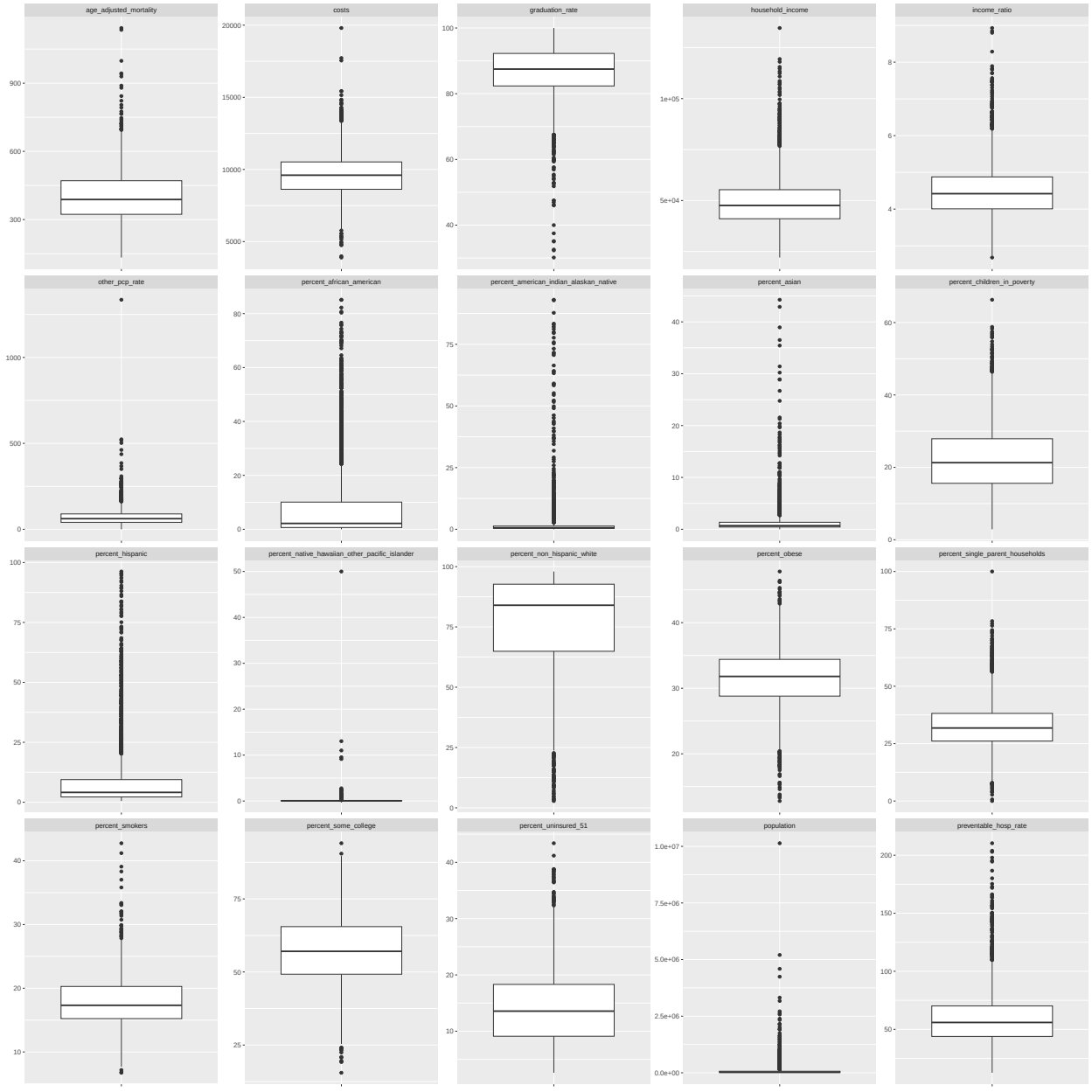
costs	other_pcp_rate	income_ratio	preventable_hosp_rate
Min. : 3896	Min. : 0.00	Min. :2.682	Min. : 12.57
1st Qu.: 8627	1st Qu.: 40.44	1st Qu.:4.007	1st Qu.: 43.91
Median : 9603	Median : 61.78	Median :4.421	Median : 55.92
Mean : 9630	Mean : 71.40	Mean :4.522	Mean : 59.88
3rd Qu.:10521	3rd Qu.: 89.34	3rd Qu.:4.877	3rd Qu.: 70.13
Max. :19803	Max. :1335.66	Max. :8.929	Max. :210.32
NA's :7	NA's :35	NA's :2	NA's :122

Table 6: Table continues below

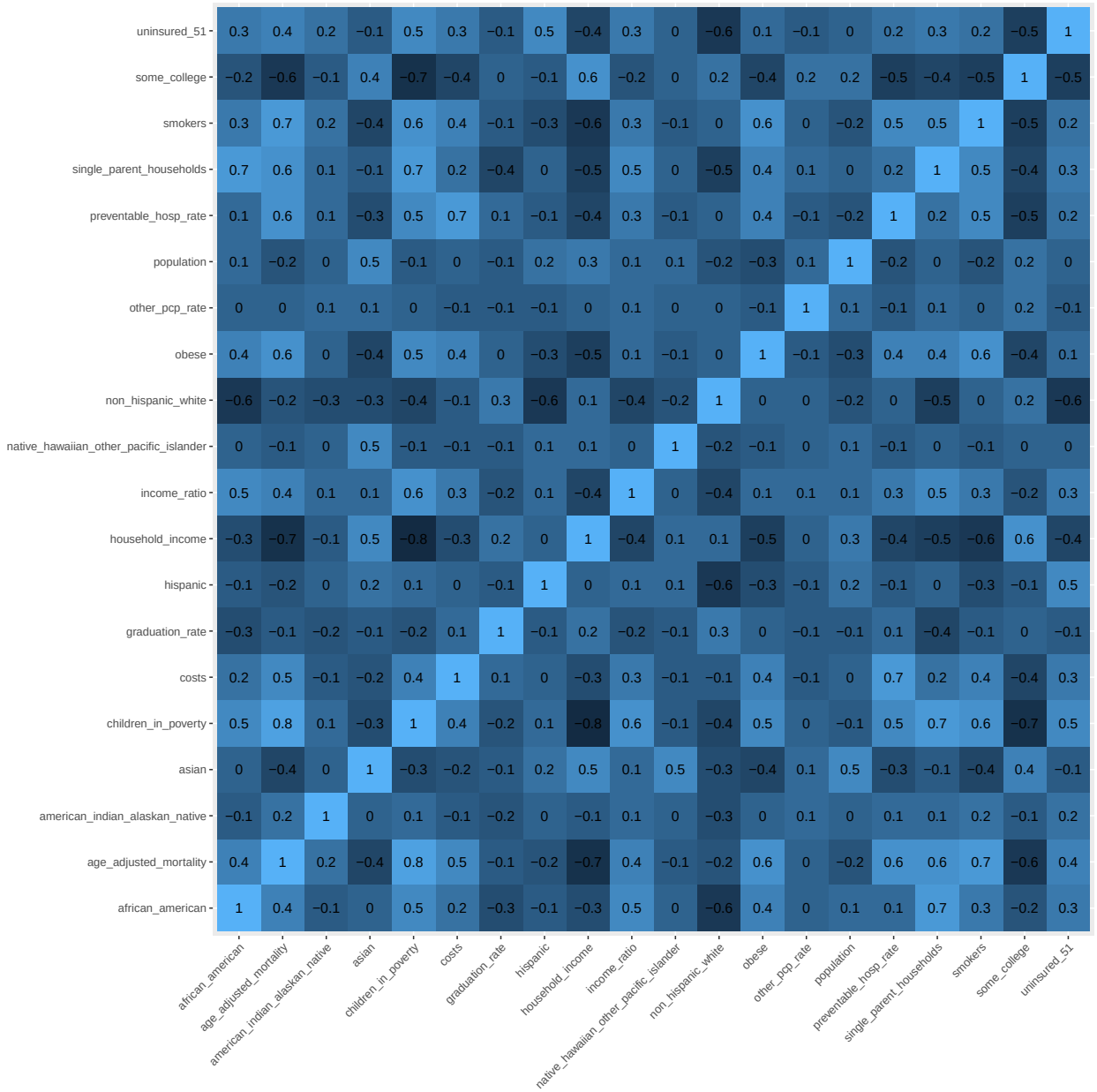
graduation_rate	percent_some_college	percent_single_parent_households
Min. : 30.14	Min. :15.51	Min. : 0.00
1st Qu.: 82.36	1st Qu.:49.24	1st Qu.: 26.14
Median : 87.50	Median :57.06	Median : 31.81
Mean : 86.19	Mean :57.23	Mean : 32.66
3rd Qu.: 92.26	3rd Qu.:65.51	3rd Qu.: 38.17
Max. :100.00	Max. :94.05	Max. :100.00
NA's :470	NA	NA's :1

percent_children_in_poverty	percent_obese	percent_smokers
Min. : 2.90	Min. :12.80	Min. : 6.735
1st Qu.:15.60	1st Qu.:28.80	1st Qu.:15.235
Median :21.30	Median :31.80	Median :17.321
Mean :22.37	Mean :31.47	Mean :17.873
3rd Qu.:27.90	3rd Qu.:34.40	3rd Qu.:20.280
Max. :66.30	Max. :47.80	Max. :42.754
NA's :1	NA	NA

2.2.1 Exploratory Analysis



2.2.2 Correlation Analysis



2.3 Model Fitting

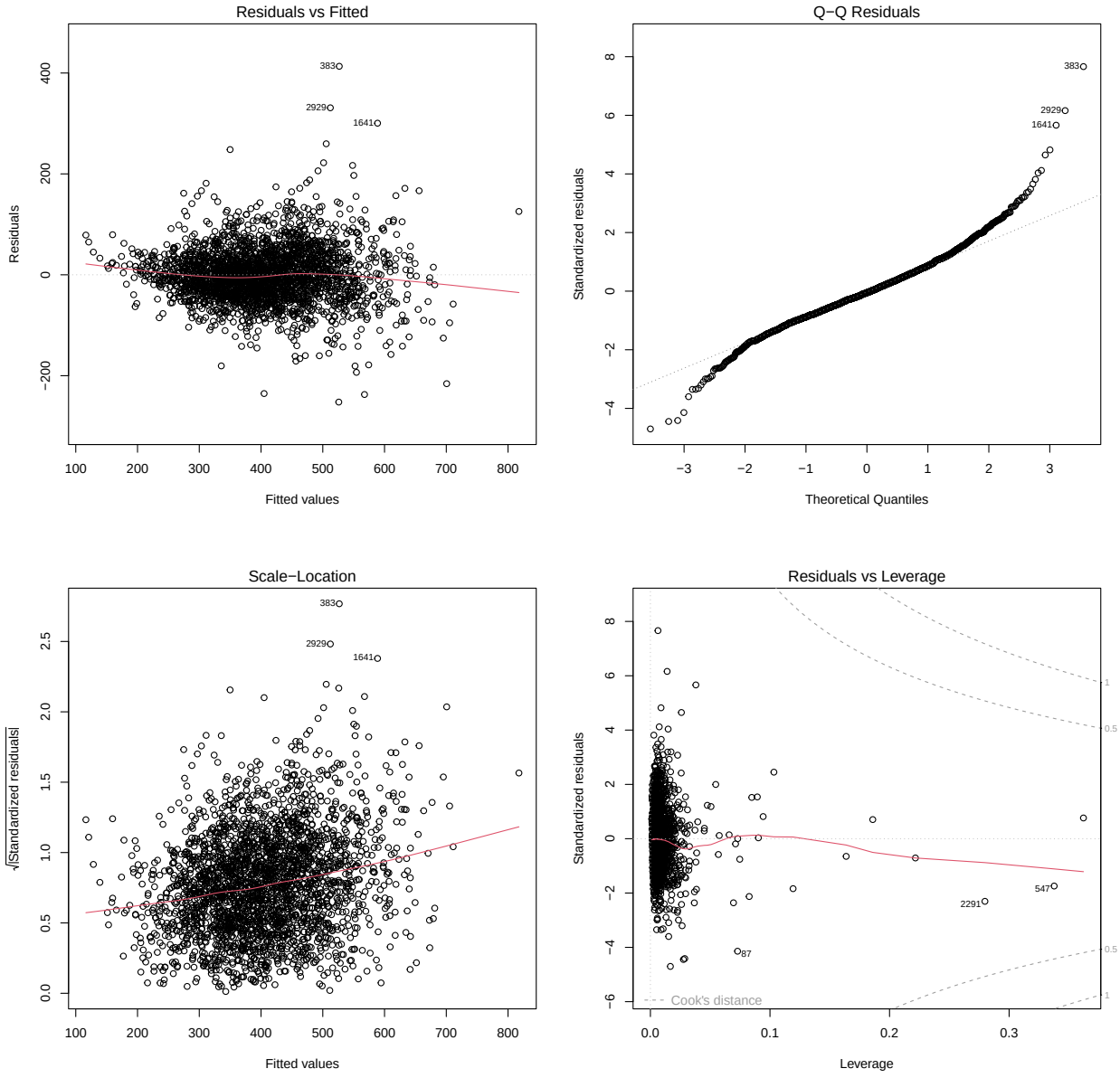
Table 8: Model Coefficients

term	estimate	p.value
(Intercept)	583.3315756	0.0000047
household_income	-0.0005119	0.0056913
population	-0.0000075	0.0368519
percent_african_american	-5.5871238	0.0000045
percent_american_indian_alaskan_native	-4.5741249	0.0003098
percent_asian	-6.7136490	0.0000029
percent_native_hawaiian_other_pacific_islander	-4.5601726	0.2536166
percent_hispanic	-6.2382143	0.0000001
percent_non_hispanic_white	-5.2450414	0.0000184
percent_uninsured_51	1.3832088	0.0000005
costs	0.0112580	0.0000000
other_pcp_rate	0.0960971	0.0000197
income_ratio	5.5144600	0.0074830
preventable_hosp_rate	0.4208420	0.0000000
graduation_rate	-0.0875965	0.5631576
percent_some_college	-0.9917633	0.0000000
percent_single_parent_households	1.2294889	0.0000000
percent_children_in_poverty	3.4167606	0.0000000
percent_obese	2.2109633	0.0000000
percent_smokers	3.7975199	0.0000000

Table 9: Model Summary

r_squared	adj_r_squared	sigma	p_value	nobs
0.7447771	0.7429113	54.11637	0	2619

2.4 Model Diagnostics



There appears no clear pattern in the `Residuals vs Fitted` plot, leading to assumption that the relation between predictor and dependent variable is linear. The `Normal Q-Q` depicts residual closely following the straight 4-degree dashed line. There is homogeneity of variance of the residuals based on the `Scale-Location`, based on the spread of the data points. There are 3 influential observations, number 87, 547 and 2291 based on `Residuals vs Leverage`.

3 Results

The summary statistics depict variables with missing data points, of which the observations were automatically excluded from the multiple linear regression model. The correlation plot depicts strong positive linear relationship between `Premature age-adjusted mortality` and `Percentage of children under age 18 in poverty` at 0.8, and negative with `Median household income` at -0.7.

Based on the model summary, the r-squared of 0.7447771 implies that the model explains 74.5 % variance in `Premature age-adjusted mortality` at the county level. All the independent variables, but `Percent Native Hawaiian other Pacific Islander` and `percent_some_college` are statistically significant at significance level $\alpha = 0.05$, implying that the significant variables are associated with the premature age mortality.

For the variables with positive coefficients, interpretations can be made that a unit increase is associated with the coefficient value increase in the premature age mortality. Similarly, for the variables with negative coefficients, interpretations can be made that a unit increase is associated with the coefficient value decrease in the premature age mortality. The analysis undertaken is limited to only the data provided from the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute and extensive analysis can be undertaken to include other variables, datasets or models of interest that may capture the influence of socio-economic attributes on premature age mortality.

4 References

Adler, N. E., & Newman, K. (2002). Socioeconomic disparities in health: Pathways and policies. Health Affairs, 21(2), 60–76. https://doi.org/10.1377/hlthaff.21.2.60

Braveman, P. A., Cubbin, C., Egerter, S., Williams, D. R., & Pamuk, E. (2010). Socioeconomic disparities in health in the United States: What the patterns tell us. American Journal of Public Health, 100(S1). https://doi.org/10.2105/ajph.2009.166082

Cheng, E., & Kindig, D. (2012). Disparities in premature mortality between high- and low-income US counties. Preventing Chronic Disease. https://doi.org/10.5888/pcd9.110120

Jobson, J. D., & Jobson, J. D. (1991). Multiple linear regression. Applied multivariate data analysis: regression and experimental design, 219-398.

Krieger, N., Chen, J. T., & Ebel, G. (1997). Can we monitor socioeconomic inequalities in health? A survey of US health departments' data collection and reporting practices. Public health reports, 112(6), 481.

Premature age-adjusted mortality*. County Health Rankings & Roadmaps. (n.d.-a). https://www.countyhealthrankings.org/health-data/health-outcomes/length-of-life/premature-age-adjusted-mortality?year=2024